PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-331355
(43)Date of publication of application : 30.11.2001

(51)Int.Cl. G06F 12/00
G06F 3/06

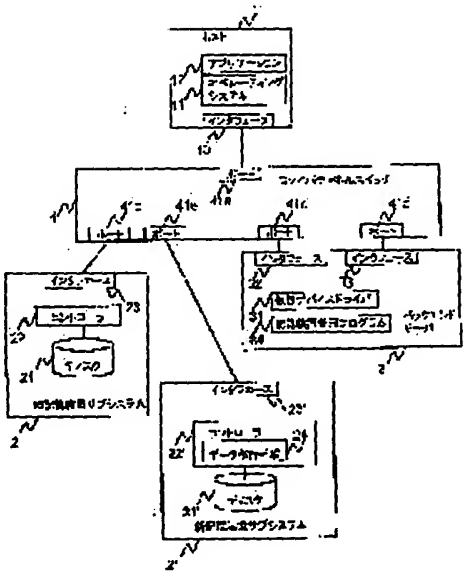
(21)Application number : 2000-152672
(22)Date of filing : 18.05.2000

(71)Applicant : HITACHI LTD
(72)Inventor : KITAMURA MANABU
ARAI HIROHARU

(54) COMPUTER SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To transmissively execute data transition among storage devices to host computers in a computer system, in which plural host computers are connected with plural storage devices.
SOLUTION: A back-end server 3 provides a host 1 having a virtual disk. The virtual disk first appears to be same as an old storage device sub-system 2 to the host 1. When data is transited from the old storage device sub-system 2 to a new storage device sub-system 2, the back-end server 3 first instructs a data transition processing to the new storage device sub-system 2, and after that, switches setting of the virtual disk and makes it correspond to the new storage device sub-system 2. Data transition among the disk devices is performed transmissively to the host 1, since this switching is transmissively executed to the host 1.

図 1



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－331355
（P2001－331355A）
(43)公開日　平成13年11月30日(2001.11.30)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
Ｇ０６Ｆ	12/00	５１４	Ｇ０６Ｆ 12/00	５１４	５Ｂ０６５
		５４５		５４５Ａ	５Ｂ０８２
	3/06	３０１	3/06	３０１Ｘ	
		３０４		３０４Ｆ	

審査請求　未請求　請求項の数２　ＯＬ　（全７頁）

(21)出願番号　特願2000－152672(P2000－152672)

(22)出願日　平成12年５月18日(2000.5.18)

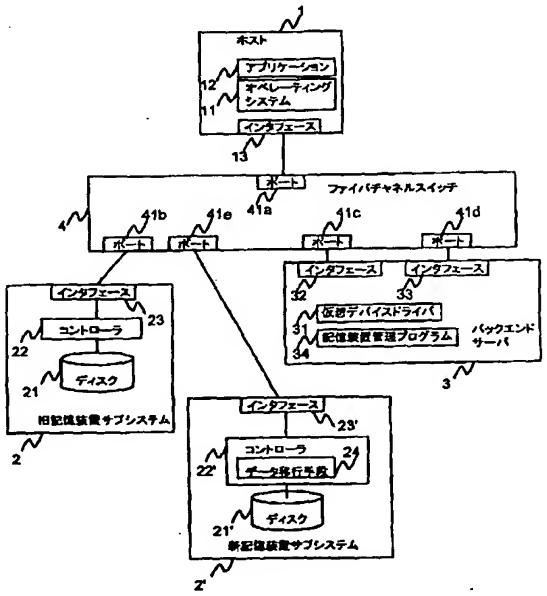
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　北村　学
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　荒井　弘治
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　100075096
弁理士　作田　康夫

Ｆターム(参考)　5B065 BA01 CE22 EA33 EA35 ZA01
5B082 FA00 HA05

(54)【発明の名称】　計算機システム

(57)【要約】
【課題】複数のホストコンピュータと複数の記憶装置が相互結合された計算機システムにおいて、記憶装置間でのデータの移動をホストコンピュータに対して透過的に実施する。
【解決手段】バックエンドサーバ３はホスト１に対し、仮想ディスクを提供する。仮想ディスクははじめ、旧記憶装置サブシステム２と同じものとしてホスト１に見える。旧記憶装置サブシステム２から新記憶装置サブシステム２にデータを移行する場合、バックエンドサーバ３は最初に新記憶装置サブシステム２にデータ移行処理を指示し、引き続き仮想ディスクの設定を切り替えて新記憶装置サブシステム２に対応させる。この切り替えはホスト１に対して透過的に実施されるため、ホスト１に対して透過的にディスク装置間のデータ移行が可能になる。

図 1



【特許請求の範囲】
【請求項１】　複数の計算機と複数の記憶装置と、前記複数の計算機と複数の記憶装置とを相互に結合するスイッチとで構成された計算機システムにおいて、該計算機システムは前記複数の計算機に対し仮想的な記憶装置を提供する手段を有し、前記仮想的な記憶装置は前記複数の記憶装置の少なくとも１つの前記記憶装置に対応する記憶装置であって、前記仮想的な記憶装置を提供する手段は、前記対応を動的に変更することを特徴とする計算機システム。
【請求項２】　請求項１における、仮想的な記憶装置を提供する手段は、前記対応を動的に変更した際に、前記複数の計算機に対しては、前記対応が変化したことを見せないことを特徴とする計算機システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、情報処理システムなどにおける記憶装置システムのデータアクセス方法に係り、特に、記憶装置内のデータ移行方法に関する。
【０００２】
【従来の技術】パソコン、ワークステーション、メインフレームなどの異なるアーキテクチャ、オペレーティングシステムを採用しているプラットフォームと複数の記憶装置とを相互に接続し、いわゆる１つのネットワークにまとめる動きが盛んになっている。これを一般に、複数の計算機をイーサネット（登録商標）（Ｅｔｈｅｒｎｅｔ（登録商標））などのネットワークで接続したLAN(Local Area Network)に対する言葉でSAN (Storage Area Network)と呼ぶ。SANは通常ファイバチャネル(Fibre Channel)という光ケーブルないし銅線の伝送路を用いて計算機と記憶装置を接続する。
【０００３】SANにはいくつかの利点があげられている。まず第１に複数の計算機から記憶装置が共通にアクセスできる環境を提供することである。第２に記憶装置同士も相互接続されることにより記憶装置間でのデータ転送が可能で、これにより、ホスト計算機に負荷をかけることなくバックアップや記憶装置間のデータコピーが実現でき、記憶装置障害時には副系の記憶装置への切り替えが可能となる。第３に、これまで個々の計算機に個々の記憶装置が接続されていたため、記憶装置の管理(装置の状態の監視、設定の変更)は接続されている個々の計算機からしかできなかったものを、特定の計算機から全ての記憶装置の管理を可能にする。また、従来のSCSI(Small Computer System Interface)では最高16台までの機器しか接続できなかったが、ファイバチャネルによって100台以上の機器をオンラインで接続でき、容易な拡張性を得られる。
【０００４】近年、SANを実現するための製品が数多く現れてきているが、実際に上記利点を生かしたものはない。とくに拡張性においては、機器のオンライン接続は物理的に可能になったものの、それを活用する基盤技術が不足している。たとえばSANにおいて、ディスク装置の交換のために新規にディスク装置を増設した場合、機器の増設はオンラインにて実施できるが、そのあとでデータの移動をユーザが明示的に行う必要がある。オンラインの機器増設でユーザがメリットを享受するには、単純なハードウェアの増設だけでなく、ハードウェアの増設に伴いデータ移動などがユーザに対して透過的に実施される必要がある。
【０００５】ディスク装置間の、オンラインのデータの移動に関しては、米国特許5680640号にその例が開示されている。米国特許5680640号はメインフレーム用ディスクを前提としたデータ移行であるが、ディスク装置間を接続する通信線を利用し、ホストとディスク装置間の接続を短時間切断するだけで、あとはユーザに透過的にディスク装置間のデータ移行を可能にしている。
【０００６】
【発明が解決しようとする課題】米国特許5680640号はユーザに対して限りなく透過的にディスク装置間でのデータ移動を可能にしている。ただし、これはメインフレーム用ディスクを前提としたデータ移行方法であり、SANにおいての適用は出来ない。米国特許5680640号では旧ディスク装置を新規ディスク装置に切り替える際、ディスク装置側の設定によって新規ディスクがあたかも旧ディスク装置であるかのようにホスト側に見せかけることが出来る。これはディスク装置のデバイス番号などの設定を操作することで可能である。
【０００７】ただし、SAN、たとえばファイバチャネル環境の場合には、個々のディスクに付与される一意なIDは、ネットワークを構成する機器（ディスク装置、ファイバチャネルスイッチ）同士のネゴシエーションによって決定され、ユーザの設定によって変えられるものではない。米国特許5680640号のデータ移行方法を用いる場合、ホストコンピュータに対して、新規ディスク装置を旧ディスク装置として見せかけることはできず、事実上ユーザに透過的なデータ移行は実現できない。本発明の目的は、ホスト、ユーザに対して透過的で、かつSANの拡張性を生かすことのできるシステムを提供することにある。
【０００８】
【課題を解決するための手段】本発明における計算機システムは、ホスト計算機、バックエンド計算機、複数の記憶装置サブシステムと、ホスト計算機とバックエンド計算機とを接続するスイッチとで構成される。ホスト計算機はバックエンド計算機を介して各記憶装置サブシステムにアクセスするが、バックエンド計算機は、ホスト計算機に対して１つないし複数の仮想的なディスク装置を提供する。ホスト計算機から仮想的なディスク装置にアクセス要求があると、バックエンド計算機では要求のあった仮想的なディスク装置の種類に応じて、実際に接

続されている記憶装置サブシステムに適宜要求を出す。

【０００９】

【発明の実施の形態】図１は、本発明を適用した計算機システムの一実施形態における構成例を示すブロック図である。計算機システムは、ホスト１、旧記憶装置サブシステム２、新記憶装置サブシステム２、バックエンドサーバ３、ファイバチャネルスイッチ４とで構成される。

【００１０】ホスト１はオペレーティングシステム１１、アプリケーション１２、インタフェース１３から構成される。オペレーティングシステム１１、アプリケーション１２は実際にはホスト１上のＣＰＵ、メモリ上で動作するが、これらハードウェアの構成要素については本発明の内容と関係が無いため省略している。実際にはホスト１以外に複数のホストコンピュータがつながる環境が一般的であるが、本発明では簡単のため、ホストコンピュータとしてホスト１のみを記載している。旧記憶装置サブシステム２はディスク２１、コントローラ２２、インタフェース２３とから構成される。ディスク２１は複数の物理ディスクをまとめて１つの論理的なディスク装置に見せかけた論理ドライブであっても、本発明の内容に変わりはない。インタフェース２３はファイバチャネルスイッチ４と接続される。新記憶装置サブシステム２も旧記憶装置サブシステム２と同様、ディスク２１、コントローラ２２、インタフェース２３とから構成される。旧記憶装置サブシステム２と異なる点は、コントローラ２２中にデータ移行手段２４が含まれることである。

【００１１】バックエンドサーバ３は仮想デバイスドライバ３１、インタフェース３２、３３から構成される。仮想デバイスドライバ３１はバックエンドサーバ３上のＣＰＵ、メモリ上で動作するソフトウェアで、ユーザによって外部から設定を変更したりあるいはプログラム自体の入れ替えをすることが可能であるが、ＣＰＵ、メモリなどのハードウェア構成要素に関しては本発明の内容と関係ないため省略している。

【００１２】ファイバチャネルスイッチ４は複数のポート４１ａ、４１ｂ、４１ｃ、４１ｄ、４１ｅ（以下総称してポート４１と略す）から構成され、ホスト１、旧記憶装置サブシステム２、新記憶装置サブシステム２、バックエンドサーバ３を相互に接続するために使用される。ポート４１ａからはいずれもポート４１ｂ、４１ｃ、４１ｄ、４１ｅにアクセスすることが可能である。そのため、ホスト１はポート４１ｂ、４１ｅから直接旧記憶装置サブシステム２や新記憶装置サブシステム２にアクセスすることもできるが、本実施形態においては、基本的にホスト１はすべてバックエンドサーバ３を介して記憶装置サブシステム２にアクセスすることとする。

【００１３】バックエンドサーバ３の役割について説明する。バックエンドサーバ３は仮想デバイスドライバ３１によって、ホスト１から見てあたかも１つないし複数のディスク装置であるかのように見える。本実施形態では、ホスト１がポート４１ｄを介してインタフェース３３を見ると、１つのディスクがつながっているように見えるものとする。以降、このディスクのことを仮想ディスクと呼ぶ。仮想デバイスドライバ３１は、最初はホスト１からは仮想ディスクが旧記憶装置サブシステム２のディスク２１と同じ物に見えるように設定されている。すなわちホスト１が仮想ディスクの論理ブロックアドレス(LBA)０あるいはLBA１にアクセスすると、仮想デバイスドライバ３１はインタフェース３２、ポート４１ｃを介して、ディスク２１のLBA０あるいはLBA１にアクセスし、結果をインタフェース３３、ポート４１ｄを介してホスト１に返す。本発明の実施形態ではインタフェース３２がディスク２１やディスク２１にアクセスするために使われ、またインタフェース３３がホスト１とのやり取りに使われるようになっているが、１つのインタフェースでこれら２つの役割を行わせることも可能である。また、仮想デバイスドライバ３１の設定を変えることで、仮想ディスクが新記憶装置サブシステム２のディスク２１に見えるようにすることも可能である。設定変更を行った場合、ホストコンピュータから見える仮想ディスクに変化はない。ファイバチャネルインタフェースを持つディスク装置の場合、ホストコンピュータからはポートIDと論理ユニット番号(LUN)で一意にディスク装置が認識できるが、仮想デバイスドライバの設定を変更して仮想ディスクがディスク２１からディスク２１に変更されたとしても、ホスト１に対して見える仮想ディスクのポートIDとLUNは変化せず、ホスト１は実際にアクセスしているディスクが変わったことの認識はない。次に新記憶装置サブシステム２のデータ移行手段２４について説明する。データ移行手段２４は米国特許5680640号に開示されているものと同様の手段を有する。データの移行が指示されると、データ移行手段２４は記憶装置サブシステム２のディスク２１の先頭から順にデータを読み出し、ディスク２１へとデータを書き込む。さらに各ブロックないしは複数ブロック単位に、データの移行が終了したかどうかを記録するテーブルを持ち、移行処理中にリードアクセスがくると、このテーブルを参照し、データ移行が済んでいない領域についてはディスク２１からデータを読み出し、データ移行が済んでいる領域についてはディスク２１のデータを返す。

【００１４】図２はデータ移行手段のもつテーブル100を表したものである。データ移行手段２４ではブロックごとにデータをディスク２１からディスク２１へとコピーしていく。テーブル１００ではそれぞれのアドレス１０１ごとにフラグ１０２を持つ。フラグ１０２が１である場合には、そのアドレスのデータはすでにディスク２１からディスク２１にコピーされたことを示し、０の場合には未コピー　であることを示す。データ移行処理

や、データ移行処理中のリード、ライト処理ではこのテーブル１００を利用する。

【００１５】図３で、データ移行手段２４の行う移行処理の流れを説明する。まずカウンタＢを用意し、初期値を０とする(ステップ２００１)。次にテーブル１００を参照し、LBA Bのフラグ１０２が１かどうかチェックする(ステップ２００２)。フラグが１の場合にはデータ移行が済んでいるため、カウンタＢを１増加する(ステップ２００５)。また、ステップ２００２で、フラグ１０２が０であれば、ディスク２１からディスク２１へとデータをコピーし(ステップ２００３)、テーブル１００の該当するフラグ１０２を１に更新し(ステップ２００４)、ステップ２００５へと進む。ステップ２００６ではディスク２１の最終LBAまで処理したかチェックする。すなわちＢがディスク２１の最終LBAを超えたかどうかチェックし、超えていれば処理を完了し、超えていなければステップ２００２に戻って、処理を繰り返す。

【００１６】次に図４で、データ移行手段が図３のデータ移行処理を行っている間に、上位ホスト、すなわち本実施形態ではバックエンドサーバ３からのライト要求があった場合の処理を説明する。この処理は簡単で、ステップ２１０１でディスク２１にデータを書き込み、ステップ２１０２でテーブル１００の該当するＬＢＡのフラグ１０２を１に更新する。つまり、データ移行処理中にライト処理が行われたＬＢＡについては、ディスク２１からのデータ移行は行われない。

【００１７】図５で、データ移行手段が図３のデータ移行処理を行っている間に、上位ホスト、すなわち本実施形態ではバックエンドサーバ３からのリード要求があった場合の処理を説明する。ステップ２２０１で、テーブル１００内のリード要求のあったＬＢＡについてフラグ１０２を参照する。ステップ２２０２でフラグ１０２が１かどうかチェックして処理を分岐する。フラグが１の場合には、そのＬＢＡについてはディスク２１からのデータ移行が完了しているので、ディスク２１からデータを読み出す(ステップ２２０３)。フラグが０の場合には、そのＬＢＡについてデータ移行が完了していないので、一旦ディスク２１からディスク２１にデータをコピーする(ステップ２２０５)。続いてテーブル１００のフラグ１０２を１に更新して(ステップ２２０６)、ステップ２２０３以降へ進む。ステップ２２０４で読み出したデータをバックエンドサーバ３に渡して処理は完了する。

【００１８】次に、本実施形態のシステムでの、旧記憶装置サブシステム２から新記憶装置サブシステム２へのデータ移行処理について、システム全体の流れを説明していく。データ移行を行う際、ユーザはバックエンドサーバ３に移行を指示する。バックエンドサーバ３から新記憶装置サブシステム２へのデータ移行処理開始の指示は、インタフェース３２を介して新記憶装置サブシステム２に伝えられる。図６はバックエンドサーバ３の処理の流れを説明している。バックエンドサーバ３は移行の指示を受けると、まず、仮想デバイスドライバ３１による仮想ディスクの動作を停止する（ステップ１００１）。これにより、仮想デバイスドライバ３１から旧記憶装置サブシステム２へのアクセスは中断され、仮想デバイスドライバ３１はホスト１から仮想ディスクに対するアクセスコマンドを受け付けても、アクセス中止が解除されるまで応答を返さない。次に記憶装置管理プログラム３４は新記憶装置サブシステム２に対してデータ移行処理の開始を指示する(ステップ１００２)。新記憶装置サブシステム２の行うデータ移行処理については後述する。ステップ１００３では、仮想デバイスドライバ３１がこれまでホスト１に見せていた仮想デバイスの設定を、ディスク２１へのアクセスを行うように変更し、ステップ１００４ではステップ１００１で中止していたアクセスを再開させる。仮想ディスクのアクセスが再開されると、ステップ１００１、ステップ１００２の間にホスト１から仮想ディスクに対してアクセスがきていた場合、そのアクセスは全てディスク２１に対して実施される。

【００１９】また、本実施形態においては、記憶装置サブシステム２とバックエンドサーバ３が直接スイッチにつながった接続形態であったが、図７のように記憶装置サブシステム２がバックエンドサーバ３を介してつながる構成であっても実現は可能である。さらに、新規に増設する記憶装置サブシステム２が図１の例のようにデータ移行手段２４をもたないような場合には、図８のようにバックエンドサーバ３側にデータ移行手段２４を持たせ、バックエンドサーバでデータ移行処理を行わせることで同様のことが実現できる。また、本実施形態ではバックエンドサーバ３を設けてホストに仮想的なディスクを見せる処理を施しているが、図９のように仮想デバイスドライバ３１、記憶装置管理プログラム３４、そしてデータ移行手段２４をファイバチャネルスイッチ４に持たせるという構成も可能である。本実施形態では、ホストコンピュータに透過的にディスク間のデータ移行ができる例を示したが、さまざまな適用先がある。仮想デバイスドライバが仮想ディスクと実際の記憶装置との対応付けを行えば、ホストコンピュータにとっては実際にデータがどの記憶装置にあってもかまわない。そのため、例えば普段は必要最低限の仮想ディスクを定義しておき、必要になった時点で動的に必要な容量の仮想ディスクを用意できるような記憶装置管理システムや、データのアクセス頻度により、ホストに透過的にデータを低速ディスクから動的に高速ディスクに移動するシステムなどに応用できる。

【００２０】

【発明の効果】本発明によれば、ホストコンピュータに対して一切透過的にディスク装置間のデータ移動や、デ

ィスク容量のオンライン拡張など、あらゆるデータ操作が可能となる。

【図面の簡単な説明】

【図１】本発明の実施形態における計算機システムの構成例を示すブロック図である。

【図２】本発明の新記憶装置サブシステムのデータ移行手段の使用するテーブルを示すテーブル構成図である。

【図３】本発明のデータ移行手段が行うデータ移行処理の流れを示すフローチャートである。

【図４】データ移行処理中にライト要求が来たときの、データ移行手段の処理の流れを示すフローチャートである。

【図５】データ移行処理中にリード要求が来たときの、データ移行手段の処理の流れを示すフローチャートである。

【図６】本発明の実施形態における計算機システムにおいて、旧記憶装置サブシステムから新記憶装置サブシステムへのデータ移行処理を行うときの、バックエンドサーバの処理の流れを示すフローチャートである。

【図７】本発明の実施形態を実現する、別の計算機システムの構成例を示すブロック図である。

【図８】本発明の実施形態を実現する、別の計算機システムの構成例を示すブロック図である。

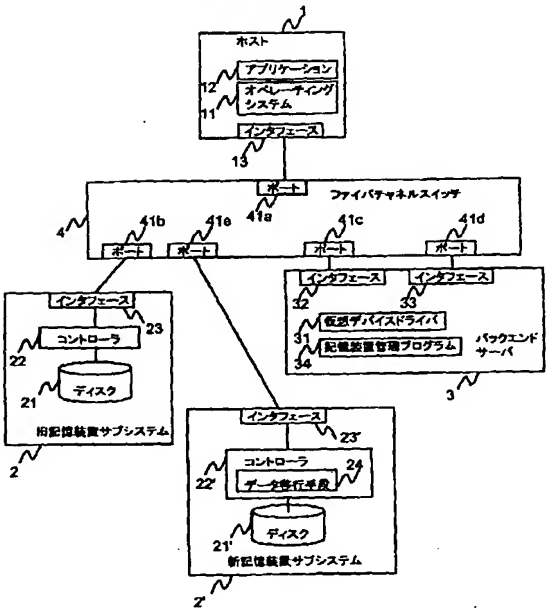
【図９】本発明の実施形態を実現する、別の計算機システムの構成例を示すブロック図である。

【符号の説明】

１…ホスト、２…旧記憶装置サブシステム、２…新記憶装置サブシステム、３…バックエンドサーバ、４…ファイバチャネルスイッチ、１１…オペレーティングシステム、１２…アプリケーション、１３…インタフェース、２１…ディスク、２２…コントローラ、２３…インタフェース、２４…データ移行手段、３１…仮想デバイスドライバ、３２…インタフェース、３３…インタフェース、３４…記憶装置管理プログラム、４１ａ…ポート、４１ｂ…ポート、４１ｃ…ポート、４１ｄ…ポート、４１ｅ…ポート。

【図１】

図 1



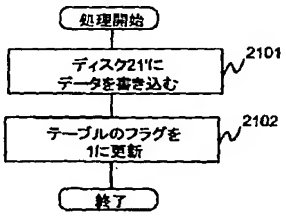
【図２】

図 2

100	101	102
アドレス	フラグ	
LBA0	1	
LBA1	1	
LBA2	0	
LBA3	0	
LBA4	1	
⋮	⋮	
LBAn	0	

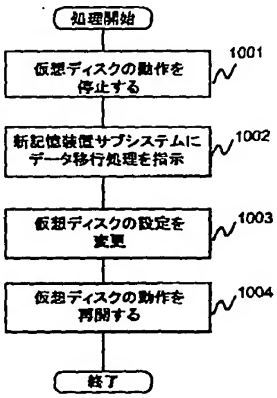
【図４】

図 4



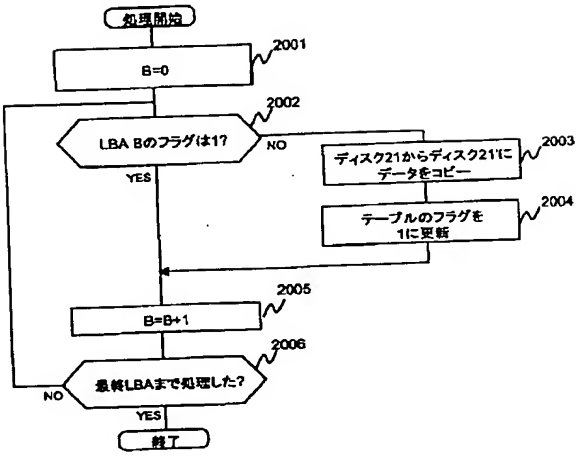
【図６】

図 6



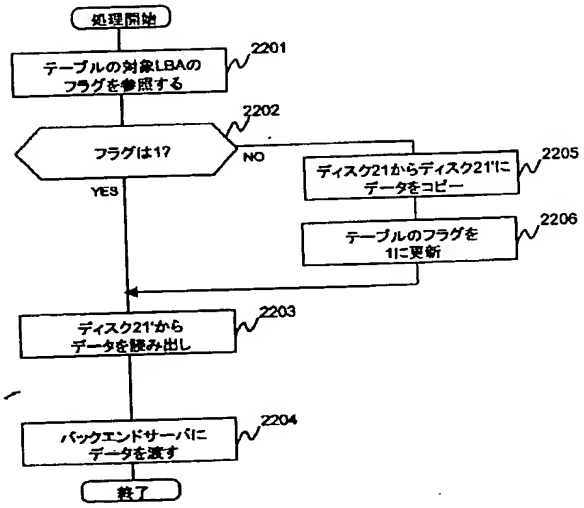
【図３】

図 3



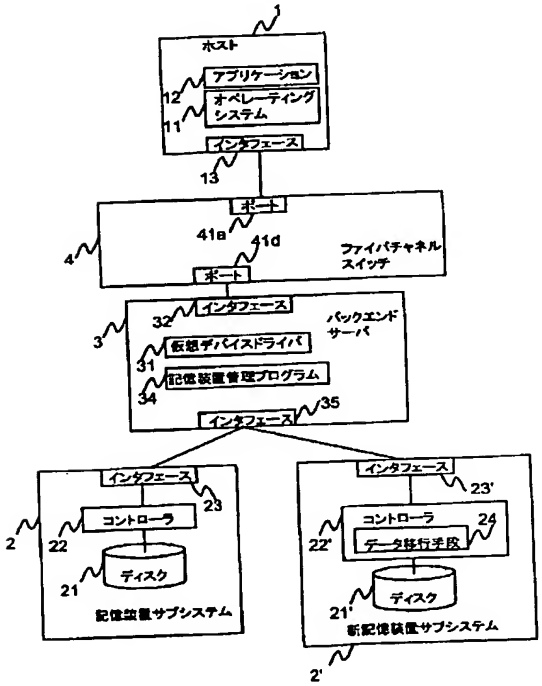
【図５】

図 5



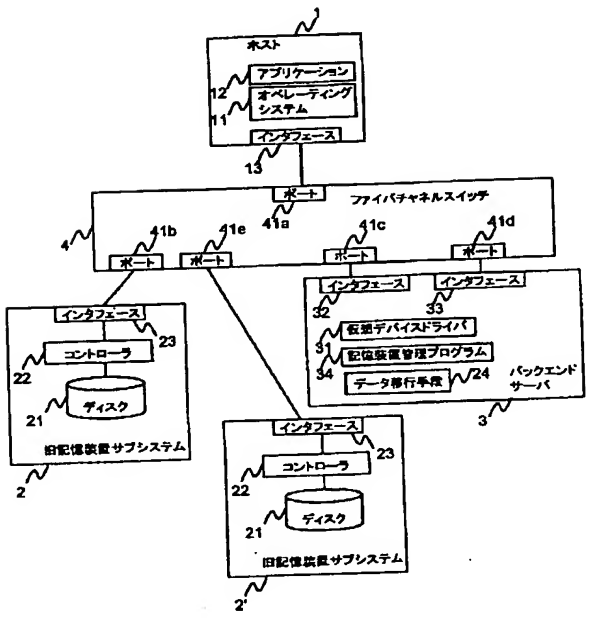
【図７】

図 7



【図８】

図 8



【図９】

図　９